DOCUMENT RESUME

ED 343 939 TM 018 105

AUTHOR Steele, Joe M.
TITLE Performance Assessment at the College Level.
PUB DATE 28 Feb 92
NOTE 9p.; Paper presented at the Annual Meeting of the Pennsylvania Educational Research Association (February 1992).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; *Educational Assessment; Feedback; General Education; Higher Education; *Knowledge Level; Literature Reviews; National Programs; *Outcomes of Education; Simulation; Standardized Tests; *Student Evaluation; Test Construction; Testing Programs; *Test Use; Thinking Skills
IDENTIFIERS Authentic Assessment; *College Outcome Measures Project; *Performance Based Evaluation

ABSTRACT

The College Outcome Measures Program (COMP) and its role in an era when standardized testing is being questioned and authentic assessment is championed are discussed. Authentic assessment should not mean discarding measurement expertise and existing technology. It is an approach to measuring the quality and level of performance that uses real settings when possible and requires the application of the kinds of skills individuals are likely to use outside the classroom. Authentic assessment plans for the measures themselves to be learning experiences that can provide feedback on processes, and it demands that measures address central and significant issues. The COMP was organized in 1976 and has developed into its present six instruments for assessing general education skills and knowledge. Four of these instruments are largely or entirely open-ended performance measures. The COMP Assessment of Reasoning and Communicating is described as an example of a performance measure consisting of six simulation activities. Elements that must be addressed in order to have confidence in conclusions drawn from outcomes data are identified, and articles based on the COMP that focus on going beyond test scores are reviewed. An eight-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

JMS: February 28, 1992
PENNSYLVANIA EDUCATIONAL RESEARCH ASSOCIATION ANNUAL MEETING

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOE M. STEELE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PERFORMANCE ASSESSMENT AT THE COLLEGE LEVEL

Joe M. Steele
College Outcome Measures Program
American College Testing Program

ED343939

There is a strong national movement that is opposed to testing, especially "standardized" testing and argues that it should be replaced with "authentic" assessment. What this group seems to want is to discard a whole technique -- multiple-choice testing -- because many such tests measure the wrong things or measure too narrow a slice of reality. But there is confusion about assessment when people begin to talk about doing away with **standardized** testing and replacing it with "authentic" testing. "Standardized" does not mean "not authentic!" The methodology of science is essential: standardization, control, reasonable levels of consistency, replicability, and explicit criteria for making judgments.

I Things Authentic Assessment requires

When I think of "authentic" assessment, I do not think of it as an approach that discards measurement expertise and technology. Instead, I think of an approach to measuring the quality and level of performance, that does five things:

First, it utilizes real settings where possible or simulates them as realistically as practicable.

Second, it requires an application of skills to the kinds of problems the individual is expected to deal with in or outside of the classroom.

Third, it assures that the measures have been validated in relevant ways, such as on effective functioning in the roles where the performance will be applied, and reasonable standards have been identified for levels of mastery.

Fourth, it plans for the measures themselves to be learning experiences which can be used to provide feedback on progress (preferably soon after the assessment) and also integrated with instruction.

Fifth, it demands that the measures address central and significant issues and are not narrowly defined, but presented as complex, ill-defined tasks that demand application of higher order thinking skills and include value clarification and ethical aspects of the situation where possible.

In order to implement such assessment, the following are needed:

a) A standardized, reliable procedure that allows comparability of performance across time and persons.
b) A fair, reasonable, and sufficiently broad range of tasks to adequately reflect the field of interest;
c) Well-defined criteria for judging levels of performance;

Measures of performance can be designed to have a narrow or broad focus. Some are designed to reflect what has been learned in specific courses and are used as evidence of the acquisition of a very specific set of skills. Sometimes performance measures are designed to represent a much broader domain of skills and

TM018105

BEST COPY AVAILABLE

knowledge. As such, they represent achievement at the much more general level of a program rather than a course. Such broadly defined measures are clearly only a sample of all the possible abilities that could be measured. They allow us to generalize about a group's level of proficiency in a whole domain of interest.

II COMP efforts to develop authentic assessments

The College Outcome Measures Program (COMP) was organized in 1976 by the American College Testing Program (ACT) to develop assessment devices and procedures to measure and evaluate the knowledge and skills (1) that undergraduate students are expected to acquire as a result of general or liberal education programs and (2) that are important to effective functioning in adult society.

In constructing instruments to assess some of the general education outcomes of college, a few fundamental assumptions were made. The first was that COMP should assess higher order cognitive skills using realistic problems rather than discipline- or content-based outcomes. Second, COMP should go beyond an assessment of academic information to assess the ability to apply skills and knowledge useful in civic, volunteer, and on-the-job roles outside of the classroom. Third, problem settings should focus on such fundamental issues as energy, pollution, waste disposal, housing, transportation, food production, health, family, schools, employment, the economy, action groups, crime, and drugs.

Working with faculty at a number of postsecondary institutions, COMP staff collected and reviewed hundreds of outcome statements for general education programs. From these efforts, six major outcomes were defined that represent some of the life-long skills that are developed in the general education part of an undergraduate degree at many institutions. Three of these outcomes are in the process-related areas of **Communicating, Solving Problems, and Clarifying Values**; three are in the content-related areas of **Functioning within Social Institutions, Using Science and Technology, and Using the Arts**. These are further defined into 35 suboutcomes. To gain efficiency in assessing these outcomes, the content and process areas are interrelated in a two-dimensional framework as shown below.

Figure 1

	Functioning within Social Institutions	Using Science and Technology	Using the Arts
Communicating	Communicating about Social Institutions	Communicating about Science and Technology	Communicating about the Arts
Solving Problems	Solving Social Problems	Solving Scientific/ Technological Problems	Solving Artistic Problems
Clarifying Values	Clarifying Social Values	Clarifying Scientific/ Technological Values	Clarifying Artistic Values

Through this framework, each content-related area is defined by the three process areas. That is, Using Science is defined as communicating about science and technology, solving scientific and technological problems and clarifying scientific and technological values. In like manner, each process-related area is defined by the three content areas. Thus, Communicating is defined as communicating about social institutions, science and technology, and the arts. The problem settings used to assess application of these skills are supported by stimulus materials currently in the adult public domain: excerpts from television documentaries, newscasts, magazine articles, newspaper advertisements, music, art, and situations encountered in civic and volunteer organizations.

The first instrument COMP developed, the Measurement Battery, took all day to administer and involved short written responses, longer essay responses and videotaped oral responses. Criterion-referenced rating scales were developed for all tasks. At that point, studies were initiated to explore the degree to which some

outcomes could be accurately assessed by multiple-choice questions rather than open-ended items. For one-third of the tasks, we found that multiple choice questions were just as effective. We also began studies to compare whether audiotaped responses could serve in lieu of videotaped responses. We determined that videotaping provided no significant additional contribution to the rating of responses. The revised instrument is thus two-thirds open-ended, with the oral response portion audiotaped in a language lab where groups could be taped at one time. This instrument, named the Composite Examination to reflect the variety of input and response formats required, can now be administered in four hours plus the 15 minute taping session. By the way, a completely multiple choice test has also been developed that does a highly accurate job of estimating what groups of students would have obtained on each scale had they taken the longer instrument. **This objective test takes half as long to administer and is the COMP instrument most institutions use for program evaluation. It is an efficient proxy measure for some assessment needs where time and resources are limited.** The paper, Review of COMP Development, describes the design and use of the Objective Test, which has been used by over 400 institutions nationally.

COMP currently supports six instruments for assessing general education skills and knowledge. Four of them are largely or entirely open-ended performance measures. A fifth instrument, the Activity Inventory, can be administered in an open-ended format. The performance measure I shall describe in a little more detail today is the COMP Assessment of Reasoning and Communicating (ARC), which was derived from the Composite Examination. It consists of six simulation activities -- three requiring longer written responses and three requiring oral responses. The three writing tasks utilize excerpts from public radio, while the three speaking tasks utilize written stimuli plus an art reproduction. Responses are couched in a role-playing context where participants **speak** to a friend, to an informal group, and at a formal meeting or **write** letters to a friend, a legislator and an administrator. Each task calls for endorsing a particular point of view and developing several specified points into a persuasive argument. The assessment can be administered in less than two hours including the taping session of oral responses.

In assessing both the spoken and written responses, equal emphasis is given to rhetorical concerns (discourse or organization of ideas), psychological concerns (making contact with and attending to the perspective of the relevant audience), and practical concerns (such as using vivid language and illustrations to dramatize and create an effect). Little emphasis is given to formal errors (such as mispronounced or misspelled words, hesitation, or punctuation) unless they noticeably detract from or obscure the message. Qualitatively defined scales at five levels of proficiency for each of three criteria result in a 36-point scale for both speaking and writing. Components of each of the six tasks call for the application of problem solving and value clarification skills that are rated by yet another set of rating scales. The evaluator must make qualitative judgments about the relevance and acceptability of each point offered to credit up to four ideas for each point addressed. These fifteen scales for the six tasks combine into a 60-point reasoning scale. Thus the assessment yields three total scores (speaking, writing, and reasoning) and nine subscores that allow some diagnostic analysis of performance.

A major element of performance assessment is the approach to defining and applying rating scales to evaluate responses. One approach involves a kind of marathon rating session in which a number of evaluators are trained to apply a holistic rating scale. Typically, this involves a reference group of example responses that represent different levels of performance. Where two raters differ in rating a response, a third person is used to resolve the difference. There are several problems with such an approach. First of all, it is not anchored by criterion-referenced scales. Therefore, whatever example responses used to define the scales may result in causing levels of performance to float from one rating session to another. This renders comparisons in performance across groups or institutions much less meaningful. Comparisons across time (to examine growth in skills) is also problematic.

An alternative approach was adopted by COMP. First, five levels of proficiency are defined for each scale. Training materials help faculty verify that they are interpreting and applying the scales in the way intended. However, for each faculty evaluator, a subgroup of participant responses are rerated by highly trained evaluators at ACT to verify reliability and level of rating. It is possible for several faculty to be in total agreement in rating a group of responses, but apply the scales in ways not intended for this instrument.

It is also possible to apply the scales reliably in the way intended but to differ in level of rating, so that responses rated A, B, C might have been rated B, C, D by ACT. Verification of both accuracy and level of rating anchor the scales so that reference group norms for freshmen and seniors can be provided and comparisons made across time and groups.

A paper on "ARC Norms and Interrelationships" provides information about technical characteristics of this instrument. You may be interested in a few relationships. In a study of 1366 seniors, the correlation between the total speaking and writing scores was .36 and correlations of these total scores with the total reasoning score were .37 for speaking and .49 for writing. Correlations with ACT, GPA, and reading scores were also at modest levels. No meaningful differences were found by age group or area of interest, although women evidenced higher means than men on all scales. Reasoning skills did not appear to be content specific; that is, science majors did not score higher in scientific reasoning than on social or artistic reasoning, nor did they score higher than social science or arts/humanities majors on scientific reasoning.

III Drawing Conclusions from Outcomes Data

Pressure to evaluate programs often seems to focus on a single, final judgment of success. There is a strong pressure to oversimplify the complex and attempt to express everything in terms of quantities -- and not just numbers, but one number: the total score. The press, legislators, and others want a scorecard that ranks institutions from best to worst. None of us is free from fault in yielding to this pressure and supplying oversimplified numbers. There is, however, an even more fundamental problem: an almost total lack of attention to the instructional program itself. We assume a single "treatment" for all students at an institution that has an effect reflected by test scores or other outcome indicators. Even if we acknowledge that different students have different experiences (courses, instructors), we assume that there is a **Program** that orchestrates these experiences in some way so that they become **a coherent instructional treatment sequenced and integrated to develop the desired skills and knowledge**. We must stop ranking schools and people and start identifying and comparing instructional programs and methods. As Paul Dressel stated in 1968:

> Effective and even inspired instruction in isolated courses does not constitute good education. A faculty member must see his/her role as one of a team composed of an entire faculty dedicated to developing a series of interrelated and coherent educational experiences which stimulate and assist the student to become a self-educating individual. (P. L. Dressel, College and University Curriculum, page x.)

Although almost forty years have past since Dressel and Mayhew's 1954 study of general education, many of their observations sound quite contemporary. Examples:

> The following two generalizations are not applicable to all teachers and classes observed but would hold for a clear majority: (1) there is more concern about the unique contribution of a particular course than there is about the combined effects of a set of general education courses or experiences; (2) integration is viewed more as a problem of organizing materials drawn from various disciplines through identification of a logical set of principles and concepts than as a characteristic of a generally educated person. (Dressel & Mayhew, General Education: Explorations in Evaluation, page 255)

> Although objective tests are heavily criticized by teachers, an objective test is more likely to be used than novel, complicated essay or rating techniques. However, the tests are often used without sufficient study or understanding or acceptance of the definitions and objectives involved, so that unfavorable results are used as a basis for discarding the test rather than as a basis for re-examination of educational practices. (Dressel & Mayhew, page 259)

A defensible assessment process must be verified before results of the assessment program itself can reasonably be expected to provide evidence for program improvement. Several critical elements must be addressed in order to have confidence in the conclusions drawn from outcomes data.

1) **Verification of program**: documentation that key instructional elements of the treatment were implemented as intended;
2) **Assurance of fit**: judgment that the chosen assessments measure or are directly related to criteria on important subsets of the intended outcomes of a program;
3) **Adequacy of sampling**: selection of representative samples of sufficient size to be meaningful;
4) **Cooperation of Participants**: adequate motivation of participants involved in the studies; and
5) **Systematic Monitoring**: completion of two or more assessments to replicate results and focus on direction and magnitude of change.

These elements constitute highly relevant evidence for funding and accrediting agencies that show that meaningful outcomes assessment is occurring. **Note that this evidence results from an evaluation of the assessment program.** If a defensible assessment process is in place, then results of the assessment program provide evidence of program effectiveness. Such judgments address the following questions:

1) Is the exiting level of participants at an acceptable level compared to programs at other institutions?

   (An answer to this question requires meaningful reference group norms by grade level, as well as comparisons with appropriate subgroups of institutions. Such norms and comparison groups require considerable time and the participation of large numbers of institutions to develop. They also must be regularly updated. In judging exiting level, the entering level of students must also be taken into account, assuring that groups compared are indeed directly comparable.)

2) Is the amount of change large enough to be practically (not merely statistically) meaningful to justify the costs of the program?

   (Sometimes it is difficult to establish how much change it is reasonable to expect for a given time period and program. Simply knowing the average amount of change other programs have shown may not adequately inform this judgment. Criterion-referencing and standard-setting by a panel of experts can assist in answering this question.)

3) What proportion of students in the program are not reaching a reasonable level of mastery?

   (Presuming the program is effective in developing relevant skills, is there a group of students for whom little growth is shown? Setting a standard for mastery could involve studies that relate performance on these skills to effective functioning in a variety of roles outside the classroom. If such a link is established, it becomes critical to assist students in achieving mastery.)

4) If changes have been introduced into the program, are they associated with increased growth and/or higher exiting levels of proficiency?

(The issue here is not to prove causality, but to find a degree of association between change in program and growth reflected by outcome indicators.)

IV Going Beyond Scores

There is general agreement these days that multiple measures are needed for assessment. Less attention has been given to the indicators of change that are derived from those measures. Means for the whole group and relevant subgroups can be monitored from year to year. Utilizing a longitudinal design, **mean gains** can be monitored for subgroups and programs from year to year. Proportions of students achieving selected levels of mastery can be monitored for relevant subgroups and programs. Yet another indicator, an achievement index, can compare performance by gender, age group and ethnic group with expected performance for national reference groups, taking into account entering achievement.

As increasing attention is given to fostering diversity and providing for the needs of many subgroups, refinements must be made in ways in which outcomes data are collected, summarized and compared. In this effort, differences in production and use of evidence must be taken into account. The following articles, based on the 15 years of experience and research in the College Outcome Measures Program (COMP), summarize problems in data collection at individual institutions, problems in design and analysis of studies to assess growth, and problems of instituting external assessments that may be examined and interpreted by audiences outside individual institutions. These are general problems, not specific to particular instruments selected for assessment.

In Distinguishing Differences, Steele (1990) describes a number of problems that can substantially affect and distort an individual institution's results from one testing or year to another. A major problem is failure to base analysis on instructionally relevant subgroups. Other problems include sampling variation and sampling bias, nonstandard testing conditions, errors in coding or reporting information, and inclusion of cases of questionable validity. The latter problem relates to adequately motivating students, including them (and faculty) as participants in the research/assessment effort, and teaching students the worth of the general education program and its evaluation.

In Improving the Accuracy of Assessments of Growth, Steele (1991) describes three approaches to estimating gains from college entry to exit. Benefits and limitations of each approach are discussed. Quasi-longitudinal studies utilizing ACT or SAT scores as an entering measure of achievement allow monitoring of both growth and exiting levels of achievement. Such studies can be quite useful to allow statewide or national comparisons of institutions of a similar type to verify maintenance of appropriate levels of performance. To guide faculty and program improvement efforts within a single institution, true longitudinal studies may be more helpful than quasi-longitudinal studies, as they are able to reflect growth in a number of areas rather than a single total score. To assist students in monitoring their growth in these skills, even more detailed and extensive tasks involving application of skills in specific course-related contexts would be required.

In Evidence for What? Fitting Assessment to the User, Steele (1991) considers judgments of program quality from the perspective of several audiences -- state and regional agencies (accrediting associations, state legislatures), college administrators and faculty, students and parents. While different users are all concerned with quality, their needs differ based on kinds of evidence required for various decisions and purposes. Presenting assessment results for different users in supportive and helpful contexts is crucial to the process. When results are misunderstood and misinterpreted, they may be used in disruptive ways rather than in support of growth and improvement. Statewide assessment efforts are reviewed from this perspective to illustrate approaches that have been more -or less- effective in encouraging program improvement.

In Using Outcomes Data to Guide Public Policy, Steele (1991) addresses limitations in utilization of data within and across institutions. Several approaches to data analysis and interpretation that make comparisons more meaningful and equitable are suggested. To monitor both program quality and diversity, two strategies are discussed. One approach provides a Growth Index of Effectiveness. A second approach is based on a Regression Index of Effectiveness.

V Assessment as Change

Meaningful outcomes assessment must be seen as a part of the instructional process. Both students and faculty must be brought to understand and trust the assessment process. For that to occur, they both must be empowered to make choices and decisions about instruction and assessment. This represents a considerable departure from current practices in assessment. It should come as no surprise that students are threatened by evaluation and are not convinced that the results will be used in positive ways, if used at all. Administrators at an institution must take on the task of proving that they are sincerely listening to students and taking action on their concerns, and they must also convince faculty to share in this endeavor.

Part of the instructional program should include explaining the rationale of both program and assessment. Both instruction and assessment should be challenging, stimulating, and productive of learning and thought. Beyond such intrinsic benefits, however, as many tangible rewards as possible should be planned. This is a strategy basic to human nature and should already be built into instruction in many ways. Requiring students to give substantial amounts of time to surveys and tests should be rewarded when possible by release from some other requirement or fee. Sometimes it may be possible to give preferential treatment in registration or eligibility. At other times, assessment activities can be seen as a direct component of a course. Sharing results with students, including group results for the institution, is an often neglected step that can increase students' interest. Faculty sometimes are at a loss to explain test results or recognize implications for program improvement based on such data. Rarely do they ask for student input and involve students in development efforts. Students represent an untapped resource in the process of change and improvement of programs.

List of References

Dressel, P. L. (1968). College and university curriculum. Berkeley, CA: McCuthan Publishing Corporation.

Dressel, P. L., & Mayhew, L. B. (1954). General education: Explorations in evaluation. Washington, DC: American Council on Education.

Steele, J. M. (1986). The Assessment of Reasoning & Communicating (ARC) Norms and Interrelationships. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).

Steele, J. M. (1989). Review of COMP Development: Instrument Design and Evaluation of the Objective Test for Use in Program Evaluation. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).

Steele, J. M. (1990). Distinguishing Differences. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).

Steele, J. M. (1991). Improving the Accuracy of Assessments of Growth. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).

Steele, J. M. (1991). Evidence for What? Fitting Assessment to the User. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).

Steele, J. M. (1991). Using Outcomes Data to Guide Public Policy. (Available from College Outcome Measures Program, American College Testing Program, Iowa City, IA).